Pair-copula constructions – an inferential perspective

Ingrid Hobæk Haff

Dissertation presented for the degree of Philosophiae Doctor (PhD)

(sfi)² Statistics for Innovation

Department of Mathematics
Faculty of Mathematics and Natural Sciences
University of Oslo
Norway

Oslo, September 2012

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo
No. 1218

ISSN 1501-7710

Cover: Inger Sandved Anfinsen.
Printed in Norway: AIT Oslo AS.

Produced in co-operation with Akademika publishing.
The thesis is produced by Unipub merely in connection with the
thesis defence. Kindly direct all inquiries regarding the thesis to the copyright
holder or the unit which grants the doctorate.

Acknowledgements

I started to work on this thesis in August 2008, funded by Statistics for Innovation, (sfi)^2. All the while, I kept a 20% position as a researcher at Norsk Regnesentral, in order to stretch the time and allow for some variation. I thank the University of Oslo, more specifically the Institute of Mathematics, for giving me this opportunity, and also Norsk Regnesentral, that has been my employer.

This has been a challenging experience, and fortunately, I have had many helpers. First, I would like thank my supervisors; I have been so lucky as to have three of them. My main supervisor, Professor Arnoldo Frigessi, has always been very enthusiastic, positive and supportive. In particular, he has helped me to believe in myself as a researcher. We have had many interesting and sometimes heated discussions, and I am very grateful for his patience with my temper. Professor Kjersti Aas, my co-supervisor and colleague at Norsk Regnesentral, is one of the most efficient and reliable persons I know. I thank her deeply for always keeping her office door open for me, and for giving me so much of her valuable time. She has also been my travel partner to a number of conferences. I have really enjoyed our trips together, and I hope that it is mutual. Both Arnoldo and Kjersti were also my co-authors on one of the papers. I acknowledge them for their contribution, and also for their important guidance throughout the work on this thesis. My second co-supervisor, Professor Ørnulf Borgan, has had a somewhat different role. I learned to know him as the teacher of a very interesting course on event history and survival analysis. I thank him for giving me the opportunity to be his teaching assistant on his statistics course for non-statisticians, three years in a row. It really has been an interesting and instructive experience.

My co-authors have not been many, but very good. I have already thanked Arnoldo and Kjersti. Another important contributor is Professor Johan Segers from l'Université Catholique de Louvain. I am grateful for our very interesting and fruitful collaboration, including a three months visit at l'UCL. While I am at it, I would like to thank all the people at l'Institut de statistique, biostatistique et sciences actuarielles for making me feel so welcome.

As mentioned earlier, I have kept my researcher position at Norsk Regnesentral part-time during these years, and that is also where I have had my

office. I thank my colleagues there for good advice, and more importantly, for an exceptionally good work environment.

These years have been important and interesting for me, but also personally difficult. I am convinced that I would not have made it without my friends and especially my closest family. My mother has been an indispensable support. I thank her for all the good advice and comfort she has given me, and for always listening to me, no matter how upset I am. Moreover, I am grateful to my father for believing in me so implicitly. He was convinced I would do a PhD long before I even started thinking about it. Finally, I want to thank my dearest sister Hilde. She is my confidant, ideal, consoler and best friend.

List of papers

This thesis consists of the following four papers.

Paper I

Hobæk Haff, I., Aas, K. and Frigessi, A. (2010): On the simplified pair-copula construction – Simply useful or too simplistic?. *Journal of Multivariate Analysis*, 101:1296–1310.

Paper II

Hobæk Haff, I. (2012): Parameter estimation for pair-copula constructions. *Bernoulli*, in press.

Paper III

Hobæk Haff, I. (2012): Comparison of estimators for pair-copula constructions. *Journal of Multivariate Analysis*, 110:91–105.

Paper IV

Hobæk Haff, I. and Segers, J.: Nonparametric estimation of pair-copula constructions with the empirical pair-copula. *Submitted for publication.*

Contents

1 Introduction

The extraordinary technological advances over the past decades have opened a world of new opportunities to statisticians. In particular, the increasing access to big data sets and growth of large databases have required the development of suited methods for multivariate analysis. This key field of statistics comprises a multitude of different models and techniques, ranging from regression models to dimension reduction by principal component or factor analysis, via networks and spatial models, such as Kriging and Markov Random Fields. Many of these assume an underlying multivariate parametric distribution.

Due to its many appealing characteristics, the multivariate normal distribution has been the, without comparison, most popular one. Though it undoubtedly has many natural areas of use, it cannot account for either heavy tails, skewness, nonlinear dependence or joint extreme events. Accordingly, many alternatives have been proposed; in particular multivariate extensions of well-known univariate distributions. The continuous ones include the multivariate Student's t and generalised hyperbolic (Barndorff-Nielsen, 1997), as well as extensions of the gamma, Pareto and many other distributions (see for instance Kotz et al. (2000)). While they are able to capture one or several of the above listed traits, they share the unfortunate feature that their flexibility decreases with the dimension, confining the range of dependence they are able to portray. Moreover, their univariate marginal distributions are all of the same type, and tend to be rather similar. For instance, in the multivariate Student's t distribution, all variables are entitled to their own location and scale parameters, and all pairs have a separate correlation, but share the degrees of freedom parameter, that governs the often crucial tails.

One manner of addressing these problems, is to transform the original variables to obtain a mathematically more tractable distribution. Using each variables own cumulative distribution function (cdf), i.e. the very natural probability integral transformation, one obtains a copula, which is precisely the theme of this thesis. More specifically, I have devoted these last years to studying a method for building flexible multivariate copulae, called pair-copula constructions (PCCs). Further, I have avoided the troublesome discrete world, thus restricting my attention to continuous distributions.

Fully aware of the many excellent introductions to the wonderful world of copulae, among those Nelsen (1999), Embrechts et al. (1999), Frees and Valdez (1998) and Genest and Favre (2007), I simply provide pieces of this vast subject, that I find particularly relevant for this thesis (Section 2). Moreover, pair-copula constructions are by no means the only manner of building multivariate copulae. I therefore consider some of the alternatives (Section 3), before I move on to my structures of choice (Section 4). Subsequently, I

summarise each of the fours papers constituting the thesis (Section 5). After that, a discussion is in order (Section 6). Finally, against Buddha's advice

> "Do not dwell in the past, do not dream of the future, concentrate the mind on the present moment",

I consider some of the choices I have made along the way, as well as possible extensions of my work (Section 7).

2 The missing link

The term "copula", derived from Latin for "link", comes from linguistics, where it denotes a word, very often a verb, that joins the subject and predicate of a sentence. In statistics, copulae may precisely be used to link variables with different margins.

That was how I was introduced to the concept. During the development of a risk management model for a bank, some of my colleagues at Norsk Regnesentral, among those my supervisor Kjersti Aas, were asked to construct a joint distribution having one beta and one log-normal margin, not unlike the example from Embrechts (2009). They came up with the following idea. Simulate from a bivariate normal distribution. Transform the samples to uniforms on $[0, 1]$ with the normal cdf. Finally, transform to beta and log-normal samples, using their respective quantile functions. A few years later, they found out that they had in fact used a Gaussian copula.

This illustrates how natural the idea is. Actually, it dates as far back as the 1940s, at least, with the work of Hoeffding (1940, 1941). Later on, Sklar introduced the name "copula". However, copulae were not straight away as fashionable as today. Genest et al. (2009a) note that their popularity increased steadily, but cautiously, from the late 1980s. The real boom came a decade later with the books of Joe (1997) and Nelsen (1999), and the introduction to finance by Frees and Valdez (1998) and Embrechts et al. (1999).

The definition of a copula is a distribution of d (≥ 2) random variables, that marginally are uniformly distributed $U[0, 1]$. According to Sklar (1959), any d-variate cdf $F_{1\ldots d}$, with univariate margins $F_1, \ldots, F_d$, may be expressed as

$$F_{1\ldots d}(x_1, \ldots, x_d) = C_{1\ldots d}\left(F_1(x_1), \ldots, F_d(x_d)\right), \tag{2.1}$$

where $C_{1\ldots d}$ is a copula, which is unique if the distribution is continuous. I have already mentioned that copulae are well suited for modelling the joint

distribution of variables with dissimilar marginal behaviour, such as a portfolio of different types of assets. That is just one of their many areas of use. In some applications, one knows the margins rather well, but has only a vague idea about the dependence between them. This is for instance the case in total risk modelling. Having found an adequate model for each of the risk types, e.g. market, credit and operational, one can link them with a copula (see for instance Aas et al. (2007)). Moreover, copulae provide a method for isolating the dependence structure from the univariate margins. Hence, they allow you to study how the variables behave jointly, while isolating their individual behaviour. As a matter of fact, many useful measures of dependence, such as Kendall's τ

$$\begin{aligned}\tau_{12} =& \mathrm{P}\left((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) > 0\right) - \mathrm{P}\left((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) < 0\right) \\ =& 4\int_0^1 \int_0^1 C_{12}(u_1, u_2) dC_{12}(u_1, u_2),\end{aligned}$$

Spearman's ρ

$$\begin{aligned}\rho_{S,12} =& 3\left(\mathrm{P}\left((X_1 - \tilde{X}_1)(X_2 - X_2') > 0\right) - \mathrm{P}\left((X_1 - \tilde{X}_1)(X_2 - X_2') < 0\right)\right) \\ =& 12\int_0^1 \int_0^1 C_{12}(u_1, u_2) du_1 du_2 - 3 = \mathrm{Cor}(F_1(X_1), F_2(X_2)),\end{aligned}$$

where $(\tilde{X}_1, \tilde{X}_2)$ and (X_1', X_2') are independent copies of (X_1, X_2), and the coefficients

$$\begin{aligned}\lambda_{U,12} =& \lim_{u \nearrow 1} \mathrm{P}\left(X_2 > F_2^{-1}(u) | X_1 > F_1^{-1}(u)\right) = \lim_{u \nearrow 1} \frac{1 - 2u + C_{12}(u, u)}{1 - u}, \\ \lambda_{L,12} =& \lim_{u \searrow 0} \mathrm{P}\left(X_2 < F_2^{-1}(u) | X_1 < F_1^{-1}(u)\right) = \lim_{u \searrow 0} \frac{C_{12}(u, u)}{u}\end{aligned}$$

of upper and lower tail dependence, respectively, rely only on the copula.

When the univariate margins are discrete, there is not just one function that satisfies (2.1), but rather a class of functions, that may be rather broad. Thus, the copula is unidentifiable. Because of that, many of the standard results and inference techniques are not valid in the discrete case. Genest and Nešlehová (2007) show that copula modelling of count data requires extra care and caution. Still they conclude that this approach may be useful also for linking discrete margins.

Despite their popularity, copulae have been severely criticised. They have even been accused of causing the financial crisis, at least the Gaussian one (Whitehouse, 2005; Salmon, 2009). The guilt of the Gaussian copula lies in its lack of tail dependence, which entails an underestimation of the joint risk of defaults. It is however unfair to blame a perfectly good model for practitioners'

abuse of it. I cannot agree with the charge "mea copula" (Monoki, C.) unless it is extended to "mea familia mea copula".

Another famous copula critic is Mikosch (2006), who compares them to the emperor's new clothes (from the tale of H.C. Andersen). Among other things, he objects to the allegedly arbitrary transformation to uniforms on $[0, 1]$, the lack of sensible copula families and the insufficient statistical theory for these models. In the discussion, Joe (2006) remarks that for some applications, other transformations may be more natural. That does not exclude the use of copulae, since the probability integral transform is an intermediate step to any other (Genest and Rémillard, 2006). As for statistical theory concerning copulae, it was already quite extensive in 2006 and has significantly evolved since then, providing the requested goodness-of-fit tests (for instance Genest et al. (2009b) and Berg (2009)) and sensitivity studies (e.g. Joe (2005) and Kim et al. (2007)) for estimation methods. The issue of copula families will be addressed later on (Sections 3 and 4). Of course, copulae do not answer all questions on dependence, but neither does any other multivariate model or method.

3 The great leap

For bivariate models ($d = 2$), there exists a long and varied list of copula families (see for instance Joe (1997)). As soon as $d \geq 3$, the catalogue of available copulae is significantly reduced (Genest et al., 2009a). Several of the well-known Archimedean copulae generalise to higher dimensions (McNeil and Nešlehová, 2009). However, they are exchangeable. Consequently, all pairwise dependencies are the same, which makes them unfit for data with more heterogeneous dependence structures. Moreover, the restrictions on their parameters, and thus on the range of dependence they can capture, become more severe with growing d. The other typical alternative in higher dimensions is an elliptical copula, most likely the Gaussian or Student's t. The former may be a good alternative when the pairwise dependencies are rather symmetric and, more importantly, extremes do not seem to occur jointly. If the data appear to be tail dependent, the Student's t copula is preferable. Though it attributes a separate correlation to all pairs of variables, these share the degrees of freedom parameter. Just like the multivariate Archimedean copulae, the Student's t is therefore best suited when the dependence between all pairs is rather similar in terms of tail behaviour.

Due to the mentioned shortcomings, Demarta and McNeil (2005) have proposed some extensions to the multivariate Student's t copula. One of those is the grouped t copula, which is derived based on the normal variance mixture

representation of the multivariate Student's t distribution. If $\boldsymbol{Z} \sim \mathcal{N}_d(\boldsymbol{0}, \boldsymbol{R})$ and $W \sim Gamma^{-1}(\nu/2, \nu/2)$, where $Gamma^{-1}$ denotes the inverse gamma distribution, are independent, then $\boldsymbol{X} = W^{1/2}\boldsymbol{Z}$ is distributed according to the d-variate Student's t distribution with scatter matrix $\boldsymbol{R}$ and ν degrees of freedom. One clearly sees that the common ν stems from W. The idea is to partition the variables into m groups, so that the dependence within each group j may be described by the same degrees of freedom ν_j. Let $V \sim U[0,1]$ be independent of $\boldsymbol{Z}$ and $W_j = G_{\nu_j}^{-1}(V)$, $j = 1, \ldots, m$, where G_ν^{-1} is the inverse cdf of the inverse gamma distribution. Further define $\boldsymbol{X} = \left(W_1^{1/2} Z_1, \ldots, W_m^{1/2} Z_d\right)$. Then the dependence structure of $\boldsymbol{X}$ is a grouped t copula. Note that by construction $W_1, \ldots, W_m$ are perfectly positively dependent, which governs the dependence between groups. Hence, not only do pairs within a specific group obey a rather similar dependence, the inter-group dependencies are quite resemblant as well.

There are also many suggested generalisations of multivariate Archimedean copulae, attempting to relax their exchangeability property. These include generalised multiplicative (Morillas, 2005; Liebscher, 2006) and hierarchical (Joe, 1997; Whelan, 2004; Savu and Trede, 2010) Archimedean copulae. The latter of these are the most flexible. They are structures consisting of $L \leq d-1$ levels of copulae, each of dimension ≥ 2. Each copula links either some of the original variables, copulae and variables or just copulae. Figure 3.1 shows an example with five variables. All the copulae must be Archimedean, but they need not be of the same type. Archimedean copulae are linked to a generator function, that must fulfil certain criteria, in particular regarding monotonicity (McNeil and Nešlehová, 2009). Hierarchical Archimedean copulae are constructed by combining the generator functions of the copulae it consists of. The resulting combined generator must satisfy the same criteria (Hofert, 2010). This imposes quite a few restrictions on the copula types that can be joined in the structures. Another result is that the parameter values must be such that the degree of dependence decreases with the level.

Hierarchical Archimedean copulae assume rather strong intra-group and considerably weaker inter-group dependence, whereas in grouped t copulae, the dependence between groups reflects the one within them. Hence, these models are particularly suited for problems with a natural, known structure, composed of homogeneous groups of variables. Otherwise, pair-copula constructions (Section 4) may be appropriate. They have the additional advantage that one does not need to know the underlying structure. Actually, I encountered these constructions in search for a model that is more flexible than the grouped t copula. The data my colleague Kjersti and I were trying to fit, were a collection of financial assets. It was natural to partition them according to type (stocks, bonds, interest rates, etc.). However, the pairwise dependencies

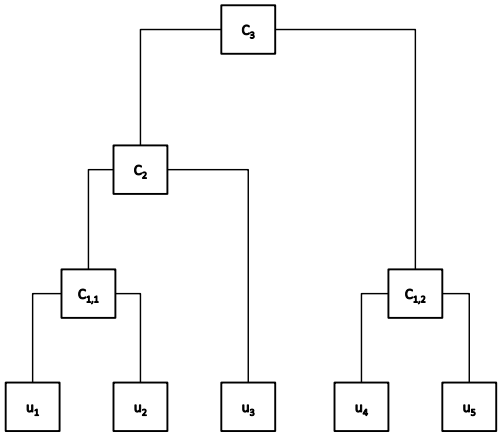


Figure 3.1: Hierarchical Archimedean copula for 5 variables.

were too heterogeneous to be well captured by the grouped t copula. Moreover, they were rather symmetric, exhibiting both upper and lower tail dependence. Hierarchical Archimedean copulae were therefore not an option for that particular problem. However, Berg and Aas (2009) compared them to pair-copula constructions in an empirical study involving a set of precipitation series and a set of equity returns. They conclude that PCCs are more suitable for high dimensional problems, due to their flexibility. Further, the authors claim that PCCs are much faster to estimate and simulate. Nonetheless, it should be noted that they did not use the more efficient algorithms proposed by McNeil (2008) or Hofert (2011).

4 Three's a crowd

In the copula world, two is company. As mentioned in the previous section, the catalogue of bivariate copulae is extensive and diversified. Moreover, these are generally more flexible than higher-dimensional counterparts. So why not build a multivariate copula based merely on bivariate ones? That is precisely the idea behind pair-copula constructions, introduced by Joe (1997).

To avoid trouble, I restrict my attention to the absolutely continuous case,

where the joint as well as all marginal probability density functions (pdfs) are defined. Further, I start with dimension $d = 3$. Let c_{123} be the copula density of the triplet (X_1, X_2, X_3), and F_i, $i = 1, 2, 3$, the corresponding marginal cdfs. To obtain the corresponding pair-copula construction, simply decompose c_{123} into the product (consult for instance Aas et al. (2009) or Paper I of this thesis to see how it is done)

$$\begin{aligned} c_{123}\left(F_1(x_1), F_2(x_2), F_3(x_3)\right) =& c_{12}\left(F_1(x_1), F_2(x_2)\right) c_{23}\left(F_2(x_2), F_3(x_3)\right) \\ & c_{13|2}\left(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2); x_2\right). \end{aligned} \tag{4.1}$$

The last factor $c_{13|2}$ is the copula density corresponding to the conditional distribution of (X_1, X_3), given $X_2 = x_2$. It is determined by

$$c_{13|2}\left(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2); x_2\right) = \frac{f_{13|2}(x_1, x_3|x_2)}{f_{1|2}(x_1|x_2) f_{3|2}(x_3|x_2)},$$

where $f_{k|3}$ and $f_{13|2}$ are the conditional pdfs of X_k, $k = 1, 3$, and (X_1, X_3), respectively, given $X_2 = x_2$. In general, the shape of this copula depends on the conditioning variable X_2, hence the notation “$; x_2$”. Also note that there are two other ways of decomposing c_{123} into a PCC.

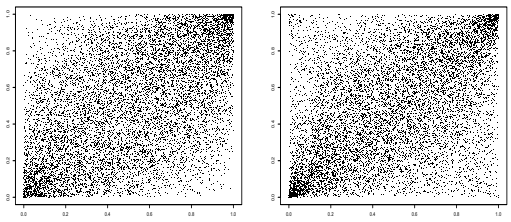
A d-dimensional copula density can be decomposed, as described above, in a product of $d(d-1)/2$ pair-copulae organised in $d-1$ levels. Each of these copulae is a function of two conditional cdfs, whose conditioning set has length 0 at the ground (first line of (4.1)) and increases by one variable with each level.

The key to these constructions is that all copulae involved in the decomposition are bivariate and can belong to different families, for instance the classical four shown in Figure 4.1, i.e. the Gaussian, Student's t, Clayton and Gumbel. There are no restrictions regarding the copula types that can be combined; the resulting structure is guaranteed to be valid anyhow. Hence, PCCs are extremely flexible and able to portray a wide range of complex dependencies (Joe et al., 2010).

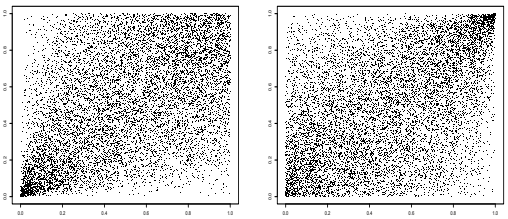
4.1 “Simplify, simplify, simplify!” (Thoreau, H. D.)

Like I mentioned above, the conditioning variables will in the general case influence the copulae constituting a PCC, not only through the pairs of arguments. In a parametric model, this means that the copula parameters are functions of them. However, as I will explain later, fast and robust inference on these structures requires the simplifying assumption that the copula shapes are constant over the values of the conditioning variables, for instance

$$c_{13|2}\left(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2); x_2\right) = c_{13|2}\left(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)\right).$$



(a) Gaussian (b) Student's t



(c) Clayton (d) Gumbel

Figure 4.1: Simulations from four different bivariate copulae.

The resulting PCC is denoted **simplified**. In Paper I, my co-authors and I explore the consequences of these constraints. We express some conditions under which a specific pair-copula decomposition of a multivariate copula is of this simplified form, in terms of standard measures of dependence. Certain copula families, such as Clayton and Student's t, turn out to fulfil the assumption, while others do not. Further, we show that the simplified PCC may be a rather good approximation, even when the simplifying assumption is far from being fulfilled. Of course, this is not always the case. But do not despair; PCCs may still be an option. You can attempt to improve the approximation by letting the copulae in the second and possibly third levels be of the general form, and use the estimation procedure proposed by Acar et al. (2012).

In simplified PCCs, the pair-copulae and conditional cdfs are linked via (Joe, 1996)

$$F_{1|2}(x_1|x_2) = \frac{\partial C_{12}(F_1(x_1), F_2(x_2))}{\partial F_2(x_2)} \tag{4.2}$$

for one conditioning variable, and in general,

$$F_{x|\boldsymbol{v}}(x|\boldsymbol{v}) = \frac{\partial C_{xv_j|\boldsymbol{v}_{-j}}(F_{x|\boldsymbol{v}_{-j}}(x|\boldsymbol{v}_{-j}), F_{v_j|\boldsymbol{v}_{-j}}(v_j|\boldsymbol{v}_{-j}))}{\partial F_{v_j|\boldsymbol{v}_{-j}}(v_j|\boldsymbol{v}_{-j})}. \tag{4.3}$$

Here, $\boldsymbol{V}$ is a random vector, not containing X, and $\boldsymbol{V}_{-j}$ is the same vector, reduced by variable V_j. The above formula enables a straightforward recursive

computation of the conditional distributions, that constitute the pair-copula arguments. However, this requires that the necessary pair-copulae $C_{xv_j|\boldsymbol{v}_{-j}}$ are present in the preceding levels, which is not true for all simplified PCCs. Consider for example the four-variate copula with density

$$\begin{aligned} c_{1234} =& c_{12}(F_1, F_2)c_{13}(F_1, F_3)c_{14}(F_1, F_4) \\ & c_{23|1}(F_{2|1}, F_{3|1})c_{34|1}(F_{3|1}, F_{4|1}) \\ & c_{34|12}(F_{3|12}, F_{4|12}). \end{aligned}$$

To compute $F_{4|12}$, you need $C_{24|1}$, but this is not one of the copulae in previous level. You may recover it by integration. However, the dimension of the required integrals increases with d. It is therefore highly preferable to construct self-contained PCCs, that have all the necessary elements available when needed.

This is guaranteed in the structures of Bedford and Cooke (2001, 2002). Their regular (R-) vines represent the d-dimensional copula with a collection of $d-1$ trees, corresponding to the PCC levels. The first of these has the d variables as nodes, and its edges stand for the pairs it connects in the ground level. Further, the edges of tree j become the nodes of tree $j+1$. Moreover, two nodes can only be connected if they fulfil the proximity condition, i.e. that the corresponding edges in the preceding tree share a node. The interested reader may consult Figure 1 in Paper III and IV for five-dimensional examples of R-vines, as well as the two subclasses, canonical (C-) and drawable (D-) vines.

4.2 All roads lead to Rome

There are numerous ways of estimating pair-copula constructions. In a parametric setting, maximum likelihood (ML) estimation is always an option, at least in theory. Moreover, since PCCs are multivariate copulae, any standard copula estimator is applicable. The most commonly used ones are the inference functions for margins (Joe, 1997, 2005), IFM, and the semiparametric (Genest et al., 1995; Shih and Louis, 1995), SP, estimators. These are two-step procedures, based on the preliminary estimation of the univariate margins, either parametrically (IFM) or nonparametrically (SP). Plugging in the resulting estimates, they maximise the log-likelihood function over all the copula parameters simultaneously. They also share the property that they require numerical optimisation.

Of course, the flexibility of PCCs has a price; they have an extensive parameter vector. The number is at least $d(d-1)/2$ if none of the pair-copulae share parameters, that is without counting the margins. It is needless to

say that optimisation over the full parameter space becomes both highly time consuming and numerically challenging with growing dimension. The ML estimator is therefore out of the question for larger models, and even the two-step procedures IFM and SP necessitate good start values to succeed.

The stepwise semiparametric (SSP) estimator is designed to exploit the level structure of PCCs, in order to speed up and facilitate the optimisation. Instead of estimating all the copula parameters at once, it proceeds level by level. Since the density of a PCC is a product of copulae, the log-likelihood function is a sum of log-copula densities, whose terms can be grouped according to the levels of the structure. The SSP procedure starts with the ground level parameters, maximising the corresponding log-likelihood terms. After that, it estimates the second level parameters in a similar manner, plugging the ground level estimates into the relevant terms, and continues level by level until it reaches the top of the structure. Moreover, the parametric univariate margins F_i, $i = 1, \ldots, d$, are substituted by their empirical counterparts $F_{in}(y) = 1/(n+1) \sum_{j=1}^{n} I(x_{ij} \leq y)$, where x_{ij}, $j = 1, \ldots, n$, are the observations of X_i. It can therefore be seen as a stepwise version of the SP estimator. An equivalent variant of the IFM estimator has also been proposed (Joe and Xu, 1996).

It is precisely the SSP estimator that has the leading role in Papers II and III. I cannot take the credit for inventing it; it was first mentioned in Aas et al. (2009). However, Paper II offers the first precise presentation of this estimator, providing both detailed estimation algorithms and large sample characteristics, such as consistency and asymptotic normality. Of course, since it performs the estimation in several steps, it has generally a larger variance than the three earlier mentioned estimators.

I have mentioned earlier that the SP estimator requires good start values. That is a perfect task for its stepwise cousin. The question is how much precision you really gain by doing a subsequent, time consuming SP estimation. In other words, is it worth it? Unfortunately, the limiting covariance matrix of the SSP estimator involves a d-dimensional integral, a property it shares with the full version, SP. Thus, computing it is in practice very difficult or even impossible. To compare these two estimators, which are the most popular for pair-copula constructions, I have therefore performed an extensive simulation study. That had the additional advantage of enabling me to explore their finite sample characteristics. This study is the theme of Paper III.

So far, I have only mentioned frequentist, likelihood based approaches. Of course, there are many other possibilities. Among the frequentist ones, you could consider method of moments type estimators, such as the ones of Clayton (1978); Oakes (1982); Genest (1987); Genest and Rivest (1993). Another alternative is the Bayesian method presented by Min and Czado (2010, 2011).

But why should we restrict ourselves to the parametric world? In Paper IV, my co-author and I propose a nonparametric estimator for simplified PCCs, that achieves the parametric convergence rate, regardless of the number of conditioning variables. This method has many potential applications, as I will discuss in the next section. Naturally, if the parametric model is correctly specified, it is less efficient than a parametric estimator. On the other hand, it is more robust.

4.3 Natural selection

The number of possible pair-copula decompositions of a multivariate copula explodes with the dimension. Even if we restrict the alternatives to R-vines, there are as many as $2^{\binom{d-2}{2}-1}d!$ different structures to choose between (Morales-Napoles, 2011), each one leading to a proper copula density. They can however be quite different in terms of computational convenience. Moreover, though it has not been completely confirmed, studies indicate that structures modelling most of the dependence in the first levels have a lower overall parameter uncertainty.

In order to be able to do inference on these constructions, one must choose the structure, i.e. the pairs to link in all levels. Further, in a parametric setting, one also has to select the pair-copula types. In the preceding section, I simply assumed that these were known. Ideally, though, one should perform all three (or possibly two) tasks simultaneously. Obviously, this is not feasible in practice.

The suboptimal solution is to do them step by step. One approach is to build the structure top down, minimising the dependence in the upper levels. Kurowicka (2011) proposed such a procedure, based on partial correlation coefficients. Another, more popular method is to start at the ground, and choose the largest possible dependencies in the first levels. The rationale is that these levels are the most influential for the total dependence. For instance, in the 50-dimensional example of Paper III, one sees that all the unconditional pairwise dependencies are well captured although the estimates in the higher levels are very unprecise. Moreover, the model uncertainty increases with the level, as I will explain later. Hence, it is preferable to model as much dependence as possible at the bottom of the structure.

Aas et al. (2009) advocate the choice of pairs with the highest tail dependence at the ground of a D-vine, while Dissmann et al. (2011) suggest an R-vine selection algorithm, based on Kendall's τ coefficients. The latter proceeds level by level, searching among all possible spanning trees, for the one with the maximum sum of absolute values of τs. At the ground, there are no restric-

tions, and the τs are estimated empirically from the data. From the second level on, the admissible trees are the ones that satisfy the proximity condition. Moreover, the copula arguments are conditional distributions. These must be computed using (4.2) or (4.3), which requires estimates of the pair-copulae in the preceding level. Hence, the algorithm involves the simultaneous selection of pair-copula types and estimation of their parameters.

The estimation is typically performed by the SSP estimator, since it must be done level by level. Further, there are several strategies for choosing the copula types. One involves the use of model selection criteria. Grønneberg (2011) shows that the use of the traditional AIC and BIC combined with semiparametric estimation is incorrect. Still, these measures appear to work rather well in practice (Dissmann et al., 2011). Nevertheless, it would be better to expand the criterion proposed by Grønneberg (2011) to PCCs. An alternative approach involves goodness-of-fit tests. At the ground level, one can for instance apply the ones studied in Genest et al. (2009b) or Berg (2009) to each of the pair-copulae. However, from the second level, the variables are no longer normalised ranks, but semiparametric estimates of conditional distributions. Consequently, the bootstrap procedures of Genest and Rémillard (2008), needed for the computation of p-values, are no longer guaranteed to be valid. Luckily, we can use the nonparametric estimator instead. As demonstrated in Paper IV, the mentioned goodness-of-fit tests with the corresponding bootstrap procedures can be applied as though the estimated conditional distributions were normalised ranks.

Because the procedures are stepwise, the selection of copula types and pairs to link at a certain level depends recursively on the copulae fitted in the lower levels. Hence the model risk grows with the level, and bad model choices will propagate throughout the structure. A more robust alternative is the structure selection procedure proposed in Paper IV, based on the nonparametric PCC estimator.

Considering that the number of PCC parameters grows fast with the dimension, it is highly commendable to reduce it somehow. One strategy is to prune the structure by identifying the pair-copulae that are not significantly different from independence. The test for conditional independence, that we propose in Paper IV, is well suited for that task. It is a Cramér-von Mises test based on the nonparametric estimator, whose test statistic is distribution free. Hence, it is easy to implement and computationally fast. A different approach is the truncation suggested by Brechmann et al. (2012).

4.4 The proof of the pudding

Since Aas et al. (2009) made pair-copula constructions more accessible, they have been applied within a number of different fields, including finance, insurance, petroleum, genetics and environmental issues.

A large portion of the publications on PCCs concern financial applications. In fact, the data example in Aas et al. (2009) involves four daily index return series. Among the many papers I could have mentioned, I have chosen two, namely Chollete et al. (2009) and Heinen and Valdesogo (2009). The former propose a multivariate model for financial time series with a regime-switching copula. They represent the univariate margins individually with a skew Student's t GARCH model. These are linked with a copula, that varies between two regimes, one corresponding to a multivariate Gaussian copula, and the other to a PCC. The latter enables them to capture the partly asymmetric dependence which is characteristic for such data. Applying the model to the stock indices of the G5 and four Latin American countries, they show its superiority compared to other popular alternatives.

Heinen and Valdesogo (2009) present an extension of the classical CAPM model for stock returns, based on a dynamic, truncated C-vine. To allow for large dimensions, their idea is to capture as much of the dependence between the stocks as possible via their dependence on a global market index, as well as on the index of the sector they belong to, just like the CAPM model. Instead of correlations, they use bivariate copulae. The ground level of the C-vine represents all pairwise relations between the global index and the sector and stock indices. All these pairs are modelled with a bivariate DCC GARCH model with either Student's t, skew Student's t or Gaussian margins. A pair-copula, whose parameters are allowed to vary over time, joins the standardised error terms. The following levels correspond to the dependencies on the sector indices, conditioning on the global index. These are modelled with static bivariate copulae. Finally, the authors link the fluctuations that cannot be explained by either the global or the sector indices, the so-called idiosyncratic variations, with a static multivariate Gaussian copula, instead of assuming them to be independent. They demonstrate the model on a set of 95 stock indices, belonging to 10 different sectors. Moreover, they compare its ability to generate in-sample estimates of the Value-at-Risk with that of the classic DCC model. Overall, the PCC based model outperforms its competitor.

Copulae have already been widely applied within insurance related problems. Pair-copula constructions are now making their entry into this field. Erhardt and Czado (2012) suggest a PCC based model for yearly claim totals from different coverage types in private health insurance. The types in question are the ambulant, in-patient and dental. For many patients, at least one

of the claim totals is zero. This means that there is a positive point mass at zero, that must be taken into account and substantially complicates matters. The univariate margins are modelled as functions of a zero claim indicator, the number of claims, given that it is larger than zero, and the claim frequency. All three of these follow a glm with covariates, such as the sex, age and zip code of the patients, with logit, log and identity links, combined with Bernoulli, zero-truncated negative binomial and log-normal distributions, respectively. The resulting claim totals from the three coverage types are joined with a PCC, whose copulae belong to either of four families, namely the Gaussian, Student's t, Clayton and Gumbel. The data used by the authors include three years of claims from a German health insurer, each year being treated separately.

Spatial models are frequently used for petroleum applications, and more generally in geostatistics. The without question most popular methods are various forms of Kriging, which assume a multivariate Gaussian distribution or copula. Kolbjørnsen and Stien (2008) present an alternative Markov model on a regular grid, with a transition kernel defined by a D-vine. More specifically, the joint pdf of the cells is decomposed into a product of conditional pdfs, where the conditioning sets are restricted to a predefined neighbourhood of conditioned cell. Each of these conditional pdfs is constructed with a D-vine. Further, the authors use a non-parametric estimator, based on bivariate Gaussian kernels. Finally, they show that method reproduces a mosaic random field rather well.

Network models have become very popular for uncovering or describing interactions between gene expressions. These networks are built by testing a sequence of conditional independence hypotheses. The state of the art is to base the building algorithm on partial correlations. Kim et al. (2011) propose a robust estimation procedure for partial correlations, based on a PCC with Gaussian copulae. They exemplify the method on expressions from eight histone genes involved in the cell cycle of yeast cells.

A consequence of the threats of global warming, is the increasing interest for renewable energy, and wind power in particular. To place the wind turbines optimally, it is important to assess the spatial dependence of wind speed. Grothe and Schnieders (2011) represent the dependence structure of daily wind velocities from weather stations all across Germany by a PCC. The margins are ARMA models with seasonal mean and volatility. An analysis of the resulting residuals, shows that the dependence is highly non-Gaussian and heterogeneous. A classic spatial model is therefore not suitable. Instead the authors build a PCC, consisting of Gaussian, Clayton, Clayton survival and Frank copulae, based on goodness-of-fit tests. To assess the optimal locations for the turbines, they maximise a set of lower tail quantiles. They conclude that the capacity should be expanded offshore, on the coast and in the South.

With a couple of co-authors, including my supervisor Arnoldo Frigessi, I am currently working on a project concerning down-scaling methods for precipitation. These methods distribute the precipitation from a global climate model, in this case ERA40, on a finer grid, according to certain criteria. This is necessary to assess the potential local impact of future climate change. The aim of our project is to compare how the different methods manage to reproduce the spatial dependence between the precipitation amounts in different grid cells, using ground data as a benchmark. For the dependence structures, we use regular vines.

5 PCC – A leitmotiv

This thesis consists of four papers. As the reader will discover, the link between them (or should I say "copula") is the pair-copula construction. The synopsis constitutes the remainder of the section.

5.1 Paper I: On the simplified pair-copula construction – Simply useful or too simplistic?

Hobæk Haff, I., Aas, K. and Frigessi, A. (2010). *Journal of Multivariate Analysis*, 101:1296–1310.

When doing inference on pair-copula constructions, the state of the art is to assume that they are of the simplified form. More specifically, this signifies that the copulae linking conditional distributions depend on the conditioning variables, merely via their arguments. As explained in Section 4.1, it is not only a convenient, but a necessary assumption for practical use, at least in the higher levels of the structure. In this paper, we present some examples of models that can be represented by a simplified PCC, and others that cannot.[1]

Further, we propose some conditions under which a specific decomposition is not of the simplified form, expressed in terms of Kendall's tau, Spearman's rho and the coefficients of tail dependence. As it turns out, these measures cannot be functions of the conditioning variables. Moreover, we provide the required form of the bivariate conditional pdfs, corresponding to the pair-

[1] In Example 4.1, we state that all elliptical distributions with a positive definite scale matrix can be represented by a simplified PCC. This is unfortunately not correct. In fact, it is only valid for the Gaussian and Student's t distributions, and more generally for the corresponding copulae. Alas, humanum est errare!

copulae, for the PCC to be simplified, if all copula arguments belong to a location-scale family, whose location and scale parameters are functions of the conditioning variables.

In practice, the true distribution will seldom be exactly a simplified PCC, but it may still be approximated by one. Therefore, we tried such an approximation on one of the example models, that is not of the simplified form. This was quite successful. The results from the true distribution and the approximated one were rather similar, even when the former was far from simplified. Hence, even if the dependence structure one is trying to model does not fulfil the assumption, it is not necessary to exclude PCCs of the simplified form. More work is however needed to understand how dense the simplified models are in the class of general PCCs.

5.2 Paper II: Parameter estimation for pair-copula constructions

Hobæk Haff, I. (2012). *Bernoulli*, in press.

There are many different estimators suited for pair-copula constructions, assuming they are of the simplified form. In this paper, I focus on some of the most popular parametric ones, including the classical maximum likelihood, the inference functions for margins and the semiparametric estimators. Their characteristics are already well-known, so they are included mostly for comparison. The stepwise semiparametric estimator, which is the last one I consider, has been suggested earlier, but has never been formally introduced. That is precisely what I do in this paper. More specifically, I present its large sample properties, and provide estimation algorithms for D- and C-vines, the two most commonly used PCC types. As one would expect, the SSP estimator is consistent and asymptotically normal. Since the limiting covariance matrix involves a multi-dimensional integral, one must resort parametric bootstrap in order to compute confidence intervals. This is explained for a precipitation data set.

Compared to the considered alternatives, the SSP estimator is, in general, asymptotically less efficient. The reason is that during the estimation at a certain level, it discards information further down in the structure about the parameters in question. Nonetheless, the loss of efficiency may be rather low, as I show in a few examples. For the set of five precipitation series, the SSP estimates are actually almost indistinguishable from the SP ones. Finally, the SSP estimator turns out to be semiparametrically efficient for the Gaussian copula.

The most attractive property of the SSP estimator, is that it, unlike its competitors, is computationally tractable even in high dimensions. Additionally, it can provide start values for the other estimators, when they are applicable. Finally, it is a natural part of parametric structure selection algorithms.

5.3 Paper III: Comparison of estimators for pair-copula constructions

Hobæk Haff, I. (2012). *Journal of Multivariate Analysis*, 110:91–105.

In Paper II, I considered four of the prevailing estimators for pair-copula constructions. Although their limiting distributions are known, the covariance matrices are difficult to compute in practice. To compare their relative performance thoroughly, I therefore had to perform an extensive simulation study, which is the theme of this paper. That also allowed me to explore finite sample characteristics.

Two of the estimators I focussed on in the previous paper, the ML and IFM estimators, rely on the parametric specification of the univariate margins. The two semiparametric ones, SP and SSP, on the other hand, use the corresponding empirical distribution functions. Since the effect of the margins on the estimation of copula parameters has already been thoroughly studied (Joe, 2005; Kim et al., 2007), I have restricted my attention to the latter two, which are also the most popular in applications. In general, the SSP estimator is asymptotically less efficient than the SP one, but on the other hand much faster, and therefore frequently applied to produce start values for the latter. Hence, it is highly relevant to assess whether a subsequent SP estimation really is worth the extra time spent.

The models in the study are D-vines. Moreover, all except one are five-dimensional. In order to explore how the type of dependence (symmetric versus asymmetric, tail dependence versus no tail dependence) affects the estimators, I have alternated between three copula families, namely the Gaussian, the Student's t and the Clayton. Moreover, I have varied the parameter values, to assess the effect of the degree of dependence. Further, I have reduced the sample size first from $n = 5000$ to 500, and then from 500 to 50. Since both estimators are based on the empirical distribution functions, they are robust towards misspecification of the univariate margins. Nevertheless, they assume that the specified dependence structure is the true one. Therefore, I have also studied how they perform when the model assumptions are not completely correct. Finally, I have explored how the stepwise estimator copes with large dimensions ($d = 50$). The full version (SP) is not included in that part, since

it is much too time consuming, and probably would be numerically unstable.

As expected, the finite sample bias and mean squared error (MSE) of the SSP estimator are generally higher than its competitor's, due to its lower asymptotic efficiency. When the degree of dependence increases, so does the difference between the two estimators, in favour of SP. This difference also augments with the level of the structure when the dependence is strong. One explanation for this is that the SSP estimator is more sensitive to the repeated transformations of the data, that follow from the recursive procedure. Overall, though, it performs rather well, compared to SP. In addition, it is consistently faster, especially for high-dimensional parameter vectors.

Obviously, the variance of both estimators increases when the sample size decreases. They manage rather well with 500 observations, but $n = 50$ is simply too small to get accurate estimates from either. Moreover, the difference between them becomes smaller. That also happens when the true model is not exactly as specified. Neither estimator is particularly robust towards misspecification of the dependence structure. However, the SP estimator appears to suffer more. Thus, SP estimation may not be worthwhile on small samples and under inaccurate model assumptions.

Up to level 20 to 30 in the 50-dimensional model, the SSP estimates are rather good. After that, however, the finite sample bias and MSE explode. The reason is that the upper levels of the structure correspond to high order dependencies, which are very difficult to estimate. Fortunately, this does really affect the corresponding lower order dependencies. The estimated distribution is in fact rather similar to the true one, even though the top level estimates are completely off. All in all, the simulation study therefore supports the use of the SSP estimator in most applications.

5.4 Paper IV: Nonparametric estimation of pair-copula constructions with the empirical pair-copula

Hobæk Haff, I. and Segers, J. *Submitted for publication.*

The estimators considered in Papers II and III are all parametric. In particular, the two semiparametric ones, SP and SSP are robust towards misspecification of the margins, but not of dependence structure, as illustrated in Paper III. Moreover, the required selection of the $d(d-1)/2$ copulae constituting the PCC is done sequentially, conditioning on choices in preceding levels. That may propagate errors throughout the structure.

In this paper, we propose an alternative, namely a nonparametric PCC

estimator, and present its asymptotic distribution at a general level of the structure. This empirical pair-copula is very similar to the classical empirical copula, only based on nonparametric estimates of conditional distributions, instead of normalised ranks. For the conditionals, we use a kind of nearest neighbour smoother, requiring a bandwidth parameter h_n. Despite that, our estimator achieves the parametric rate of convergence, regardless of the number of conditioning variables, thanks to the simplifying assumption, described in Paper I.

To substantiate our conjecture on the estimator's asymptotics, we have conducted a simulation study. The results certainly support our allegations. Choosing a sensible bandwidth h_n is however important. Since the smoothing is done on the uniform $(0, 1)$ scale, this parameter does not depend on the margins, but only on the dependence structure. As it turns out, it is advisable to undersmooth. Actually, the value $0.5n^{-1/3}$ appears to work rather well overall.

Further, we propose a resampling scheme, which is required for estimating confidence intervals or computing critical values for hypothesis tests. The approach, inspired by the multiplier bootstrap of Rémillard and Scaillet (2009), is fast and easy to implement. We have tested the procedure, and compared it to parametric alternatives. Obviously, a parametric estimator is more efficient under correct model assumptions. The nonparametric method is, on the other hand, more robust.

The empirical pair-copula has a number of potential applications. Among others, we can construct nonparametric estimators of dependence measures, such as conditional Spearman rank correlations. Actually, we suggest a selection algorithm for regular vines based on these estimates, highly influenced by the procedure of Dissmann et al. (2011). The differences are that we estimate the copulae and conditional distributions nonparametrically, and that we use Spearman's ρ as a measure of dependence instead of Kendall's τ. As these two measures quantify the same type of dependence, the substitution should not influence the results too much. Hence, when the parametric model is well specified, we expect the two algorithms to select virtually the same structure, as shown in an example involving nine financial series. Further, the empirical pair-copula can be applied in tests for conditional independence, aiming at pruning the structure. We propose a Cramér-von Mises test, whose statistic is distribution free. Hence, critical values are fast and straightforward to compute. Again, simulations indicate that the test works well. We also demonstrate its use on a set of five series of daily precipitations. Finally, we suggest a goodness-of-fit test, which may facilitate the selection of parametric copulae in a given structure. The test in question is again a Cramér-von Mises test. It is actually the one of Genest and Rémillard (2008), where we adopt

the same bootstrap procedure for computing critical values, simply replacing normalised ranks by estimated conditional distributions. This approach is also supported by simulations.

6 Closing

The fewer restrictions a model imposes, the more flexible it becomes. However, this usually also implies more complexity, for instance a greater number of parameters. The PCC is a very malleable model, that can portray a wide range of dependencies. Obviously, this has a cost, both in terms of computational effort and model uncertainty. Therefore a PCC should only be applied on problems that call for such flexibility. If the dependence structure in question is rather homogeneous, at least when split into groups, then a multivariate elliptical or Archimedean copula, possibly a grouped or hierarchical version thereof, may be a more sensible model. And why not consider the multivariate normal distribution, if it seems suitable? The moral is: do not crack a nut with a sledgehammer. Still, if one is not certain that a simpler model is adequate, it is advisable to at least compare it with a PCC, to check that the results from the two really are similar.

In its most general form, the PCC is very little restrictive, but that much less useful, since inference is virtually impossible. Without the regular vine building algorithm, the computational effort becomes too large, even in lower dimensions. Another central assumption is the simplifying one, which is essential to virtually all inference methods for PCCs. For instance, the parametric convergence rate of the nonparametric estimator is only obtained under this assumption. Even when the true distribution is far from simplified itself, such a PCC may serve as a good approximation. Otherwise, the approximation may be improved by relaxing this condition for the copulae at the second level, using the estimator of Acar et al. (2012). In any case, it is a good idea to apply their method as a diagnostic, to test whether the assumption is reasonable.

When choosing an estimator for PCCs, one is generally faced with conflicting concerns, for instance precision on one hand versus computing time and convenience on the other. The stepwise semiparametric estimator appears to be a good tradeoff between the two, in most applications. Overall, it performs rather well. It is very fast and moreover numerically stable, as long as the amount of data is sufficient. Further, it copes with high dimensions, as opposed to most other relevant estimators. However, the performance of SSP relative to SP, reduces when the dependence becomes strong. In such cases, it is advisable, whenever possible, to use the full version, with SSP estimates as start values. If the sample size is very low in view of the dimension, e.g.

$n = 50$ for $d = 5$, both the SP and the SSP estimators perform poorly. ML and IFM are likely to get the same problems. Further, there is no reason to believe that the nonparametric estimator will behave any better; quite the contrary, since it imposes fewer restrictions on the copulae to estimate. In such cases, a Bayesian estimator, preferably with a prior that is not too vague, may be a better option. I would also recommend to consider a less complex model.

Another potential clash is the one between efficiency and model uncertainty. Selecting both the structure and copula types is non-trivial. The state of the art is to use stepwise, parametrically based algorithms. These suffer under a model uncertainty that increases with the level and a potential propagation of errors throughout the structure. The nonparametric estimator may serve as a remedy for the latter, since it does not rely on a parametric specification. Then again, parametric estimators are asymptotically more efficient when the assumed model is reasonable. One alternative is to use the two in combination. Choose the vine structure and copula types with the nonparametric estimator. Afterwards, estimate the copula parameters with a suitable parametric estimator, for instance SP or SSP. This procedure may reduce the model risk, which is important since the parametric estimators generally are not robust towards misspecifications of the model. An extra step in the nonparametric selection algorithm, could be the detection of independence copulae. By pruning the structure, one reduces the parameter vector, and consequently the model complexity.

Most applications of PCCs involve data that are dependent not only across variables, but also in time. That violates the assumption of independent, identically distributed observations, which is made by virtually all inference methods for PCCs. The standard procedure is to apply some time series model to the data, e.g. ARMA or GARCH, and model the residuals with a PCC. This is a useful approach, but perhaps not entirely satisfying. It will always be asymptotically less efficient than estimating the entire model simultaneously, and may lead to an underestimation of the dependence parameter uncertainty. Chen and Fan (2006) showed that, under certain conditions, the properties of the SP estimator are unaffected by the preliminary time series filtration step, and Rémillard (2010) did likewise for goodness-of-fit tests based on the empirical copula. Corresponding results for the SSP estimator are however not established. That is certainly an important subject for future research, considering the extensive use of this estimator. Alternatively, one could develop an extended, dynamic PCC model, such as the model of Heinen and Valdesogo (2009), that takes into account not only the dependence between variables observed simultaneously, but also across time.

7 Backing into the future

It is always easy to be wise after the fact. As Thoreau says

> "Never look back unless you are planning to go that way".

Still, I would like to dwell a little longer on the choices I have made during this almost four year long journey.

The first thing that I would like to point out, is that the definition of a pair-copula construction is not completely consistent. In Paper I, we denote the full joint distribution a PCC, whereas as in the remaining papers, this term is reserved for the dependence structure. The latter is in my opinion more correct. It was however convenient to include the margins in Paper I, since we were particularly interested in the characteristics of conditional distributions of the original variables. In any case, we left more questions unanswered in that paper than we would have liked. For instance, how does one know, or at least test, whether the simplifying assumption is reasonable? And if one decomposition is not of the simplified form, does there exist an alternative one that is? The first question is partly treated by others (Acar et al., 2012), whereas the last one remains to be investigated.

Paper II was, to be quite honest, a lot of hard work. Despite the huge improvements due to the excellent comments and suggestions of the associate editor and two referees, it still reveals my lack of experience for writing such papers. More specifically, I am thinking about the incredibly cumbersome notation and the somewhat clumsy proof in the appendix. Naturally, this paper also leaves a few loose ends. For convenience, I only considered the sub-categories C- and D-vines. Obtaining equivalent results for more general regular vines is however straightforward. More importantly, one should study the effects of a preliminary time series filtration of the data, as discussed in Section 6.

Generalising the results from limited cases in a simulation study feels somewhat risky. The choices I made in Paper III seemed reasonable at the time, but I cannot exclude that other choices might have lead to at least partly different conclusions. Nevertheless, I think my simulation study can be useful for practitioners and a starting point for future investigations. Again, I have restricted my attention to D-vines. Though I believe the results would be similar for C- and more general R-vines, this remains to be tested. Moreover, I do not provide any alternatives for small sample sizes. I leave that responsibility to others.

The procedures suggested in Paper IV are mostly fast and easy to implement. An exception is the goodness-of-fit test, which is nowhere near fast enough for high-dimensional PCCs. If the influence function of the parametric estimator in question is known and possible to compute, a multiplier bootstrap approach could be an option. Otherwise, it is necessary to come up with some smart way of speeding up the current routine. Moreover, we do not provide a procedure for selecting the bandwidth parameter. Though the constant value we have used appears to work rather well overall, one should investigate whether there is a more optimal method for choosing it in each case. Finally, our original plan was to write a more theoretical paper, with formal proofs of the asymptotics. Unfortunately, that turned out to be much more difficult than anticipated. Therefore, we contented ourselves with an "AOK proof" (to cite Andrew Patton's presentation at the copula conference in Montréal in June last year), based on simulation. We have however not quite given up the hope of providing the proof sometime in the future.

References

Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44:202–209.

Aas, K., Dimakos, X., and Øksendal, A. (2007). Risk capital aggregation. *Risk Management*, 9:82–107.

Acar, E., Genest, C., and Nešlehová, J. (2012). Beyond simplified pair-copula constructions. *Journal of Multivariate Analysis*, 106. in press.

Barndorff-Nielsen, O. (1997). Normal inverse Gaussian distributions and stochastic volatility modelling. *Scandinavian Journal of Statistics*, 24:1–13.

Bedford, T. and Cooke, R. (2001). Probabilistic density decomposition for conditionally dependent random variables modeled by vines. *Annals of mathematics and Artificial Intelligence*, 32:245–268.

Bedford, T. and Cooke, R. (2002). Vines – a new graphical model for dependent random variables. *Annals of Statistics*, 30:1031–1068.

Berg, D. (2009). Copula goodness-of-fit testing: an overview and power comparison. *European Journal of Finance*, 15:675–701.

Berg, D. and Aas, K. (2009). Models for construction of multivariate dependence. *European Journal of Finance*, 15:639–659.

Brechmann, E. C., Czado, C., and Aas, K. (2012). Truncated regular vines in high dimensions with application to financial data. *Canadian Journal of Statistics*, 40:68–85.

Chen, X. and Fan, Y. (2006). Estimation of copula-based semiparametric time series models. *Journal of Econometrics*, 130:307–335.

Chollete, L., Heinen, A., and Valdesogo, A. (2009). Modeling international financial returns with a multivariate regime-switching copula. *Journal of Financial Econometrics*, 7:437–480.

Clayton, D. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65:141–151.

Demarta, S. and McNeil, A. (2005). The t copula and related copulas. *International Statistical Review*, 73:111–129.

Dissmann, J., Brechmann, E., Czado, C., and Kurowicka, K. (2011). Selecting and estimating regular vine copulae and application to financial returns. Submitted for publication.

Embrechts, P. (2009). Copulas: A personal view. *Journal of Risk and Insurance*, 76:639–650.

Embrechts, P., McNeil, A. J., and Straumann, D. (1999). Correlation: Pitfalls and alternatives. *Risk*, 12:69–71.

Erhardt, V. and Czado, C. (2012). Modeling dependent yearly claim totals including zero claims in private health insurance. *Scandinavian Actuarial Journal*. DOI: 10.1080/03461238.2010.489762.

Frees, E. and Valdez, E. (1998). Understanding relationships using copulas. *North American Actuarial Journal*, 2:1–25.

Genest, C. (1987). Frank's family of bivariate distributions. *Biometrika*, 74:549–555.

Genest, C. and Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12:347–368.

Genest, C., Gerber, H. U., Goovaerts, M. J., and Laeven, R. J. A. (2009a). Editorial to the special issue on modeling and measurement of multivariate risk in insurance and finance. *Insurance: Mathematics and Economics*, 44:143–145.

Genest, C., Ghoudi, K., and Rivest, L. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82:543–552.

Genest, C. and Nešlehová, J. (2007). A primer on copulas for count data. *ASTIN Bulletin*, 37:475–515.

Genest, C. and Rémillard (2008). Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. *Annales de I'Institut Henri Poincaré. Probabilités et Statistiques*, 44:1096–1127.

Genest, C. and Rémillard, B. (2006). Discussion of "Copulas: Tales and facts". *Extremes*, 9:27–36.

Genest, C., Rémillard, B., and Beaudoin, D. (2009b). Goodness-of-fit tests for copulas: a review and power study. *Insurance: Mathematics and Economics*, 44:199–213.

Genest, C. and Rivest, L.-P. (1993). Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American Statistical Association*, 88:1034–1043.

Grønneberg, S. (2011). The copula information criterion and its implications for the maximum pseudo-likelihood estimator. In *Dependence Modeling: Vine Copula Handbook*, pages 113–138. World Scientific Publishing Co.

Grothe, O. and Schnieders, J. (2011). Spatial dependence in wind and optimal wind power allocation: a copula based analysis. *Energy Policy*, 39:4742–4754.

Heinen, A. and Valdesogo, A. (2009). Asymmetric CAPM dependence for large dimensions: the canonical vine autoregressive model. ECORE discussion paper 101, CORE, Université Catholique de Louvain.

Hoeffding, W. (1940). Scale-invariant correlation theory. In Fisher, N. and Sen, P., editors, *The Collected Works of Wassily Hoeffding*, pages 57–107. Springer-Verlag, New York.

Hoeffding, W. (1941). Scale-invariant correlation measures for discontinuous distributions. In Fisher, N. and Sen, P., editors, *The Collected Works of Wassily Hoeffding*, pages 57–107. Springer-Verlag, New York.

Hofert, M. (2010). *Sampling Nested Archimedean Copulas with Applications to CDO Pricing*. PhD thesis, Universität Ulm.

Hofert, M. (2011). Efficiently sampling nested Archimedean copulas. *Computational Statistics & Data Analysis*, 55:57–70.

Joe, H. (1996). Families of m-variate distributions with given margins and m(m-1)/2 dependence parameters. In *Distributions with Fixed Marginals and Related Topics*, pages 120–141. IMS, Hayward, CA.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall, London.

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94:401–419.

Joe, H. (2006). Discussion of "Copulas: Tales and facts". *Extremes*, 9:37–42.

Joe, H., Li, H., and Nikoloulopoulos, A., K. (2010). Tail dependence functions and vine copulas. *Journal of Multivariate Analysis*, 101:252–270.

Joe, H. and Xu, J. (1996). The estimation method of inference functions for margins for multivariate models. Technical Report 166, University of British Columbia, Department of Statistics.

Kim, G., Silvapulle, M., and Silvapulle, P. (2007). Comparison of semiparametric and parametric models for estimating copulas. *Computational Statistics and Data Analysis*, 51:2836–2850.

Kim, J.-M., Jung, Y.-S., Choi, T., and Sungur, E. (2011). Partial correlation with copula modelling. *Computational Statistics and Data Analysis*, 55:1357–1366.

Kolbjørnsen, O. and Stien, M. (2008). The D-vine creation of non-Gaussian random fields. In *Procedings of the Eight International Geostatistics Congress*, pages 399–408. GECAMIN Ltd.

Kotz, S., Balakrishnan, N., and Johnson, N. (2000). *Multivariate Continuous Distributions*. Wiley, 2nd edition edition.

Kurowicka, D. (2011). Optimal truncation of vines. In *Dependence Modeling: Vine Copula Handbook*, pages 233–248. World Scientific Publishing Co.

Liebscher, E. (2006). Modelling and estimation of multivariate copulas. Working paper, University of Applied Sciences Merseburg.

McNeil, A. (2008). Sampling nested Archimedean copulas. *Journal of Statistical Computation and Simulation*, 78:567–581.

McNeil, A. and Nešlehová, J. (2009). Multivariate Archimedean copulas, d-monotone functions and l_1-norm symmetric distributions. *The Annals of Statistics*, 37:3059–3097.

Mikosch, T. (2006). Copulas: Tales and facts. *Extremes*, 9:3–20.

Min, A. and Czado, C. (2010). Bayesian inference for multivariate copulas using pair-copula constructions. *Journal of Financial Econometrics*, 8:511–546.

Min, A. and Czado, C. (2011). Bayesian model selection for multivariate copulas using pair-copula constructions. *Canadian Journal of Statistics*, 39:239–258.

Morales-Napoles, O. (2011). Counting vines. In *Dependence Modeling: Vine Copula Handbook*, pages 189–218. World Scientific Publishing Co.

Morillas, P. (2005). A method to obtain new copulas from a given one. *Metrika*, 61:169–184.

Nelsen, R. (1999). *An Introduction to Copulas*, volume 139 of *Lecture notes in Statistics*. Springer Verlag, New York.

Oakes, D. (1982). A model for association in bivariate survival data. *Journal of the Royal Statistical Society, Series B*, 44:414–422.

Rémillard, B. (2010). Goodness-of-fit tests for copulas of multivariate time series. Working paper series, HEC Montreal. Available at SSRN: http://ssrn.com/abstract=1729982.

Rémillard, B. and Scaillet, O. (2009). Testing for equality between two copulas. *Journal of Multivariate Analysis*, 100:377–386.

Salmon, F. (2009). Recipe for disaster: The formula that killed wall street. *Wired Magazine*, March 17.

Savu, C. and Trede, M. (2010). Hierarchies of Archimedean copulas. *Quantitative Finance*, 10:295–304.

Shih, J. and Louis, T. (1995). Inferences on the association parameter in copula models for survival data. *Biometrics*, 51:1384–1399.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Stat. Univ. Paris*, 8.

Whelan, N. (2004). Sampling from Archimedean copulas. *Quantitative Finance*, 4:339–352.

Whitehouse, M. (2005). How a formula ignited market that burned some big investors. *Wall Street Journal*, September 12.

I

II

III

IV